DOCUMENT RESUME

ED 453 230 TM 032 581

AUTHOR Xie, Yiyu
TITLE Dimensionality, Dependence, or Both? An Application of the Item Bundle Model to Multidimensional Data.
PUB DATE 2001-04-00
NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Middle School Students; Middle Schools; *Sciences; Test Construction
IDENTIFIERS *Dimensionality (Tests); *Item Bundles; Item Dependency Model

ABSTRACT

A data set collected by the Science Education for Public Understanding Program (SEPUP) that has both dimensionality and dependence characteristics was studied, and an item bundle model nested in a multidimensional random coefficients multinomial logit (MRCML) model (R. Adams, W. Wang, and M. Wilson, 1997) was applied to this data. Data are from the field test of an assessment system for a yearlong middle school science curriculum, "Issues, Evidence, and You" (IEY). The SEPUP link tests have both multidimensionality and item dependence issues. Because the performance assessment was time consuming, only a small number of items could be given to each student during the testing period. For this reason, each item was specifically designed to be multidimensional and scored on a number of variables or elements. A multidimensional item bundle analysis suggests that most of the items are dependent within each bundle, no matter whether the bundles are unidimensional or multidimensional. Analyses showed that interaction effects did exist for most of the pairwise score combinations. The effects were more prominent for the bundles in which items measure content knowledge. Taking the item dependence into account may make the correlation among latent dimensions more accurate and meaningful. Nine appendixes contain additional information about the analyses. (Contains 18 references.) (SLD)

Dimensionality, Dependence, or Both?

An Application of the Item Bundle Model to Multidimensional Data

Yiyu Xie

Policy, Organization, Measurement & Evaluation
Graduate School of Education
University of California at Berkeley

AERA, 2001

INTRODUCTION

Over the last two decades, item response modeling has been widely used in educational measurement research as well as by many large test publishers. An item response model is a mathematical model that defines a relationship between the observed examinee test performance and the unobserved traits or abilities assumed to underlie performance on the test. Like any mathematical model, it has a set of assumptions. The two basic assumptions are dimensionality and local independence.

Item response models that assume a single latent trait or ability are referred to as unidimensional. In unidimensional item response models, it is assumed that only one latent trait or ability is necessary to account for examinee test performance. In reality, this assumption is extremely difficult to meet, because there are often other cognitive, personality, and test-taking factors that might influence test performance. Therefore, instead of strictly assuming one latent trait is being measured, it is usually assumed that the test items are measuring a dominant component or factor that underlies performance on the test. Models that assume more than one ability are necessary to account for examinee test performance are referred to as multidimensional. A set of test items can be constructed to measure a set of D latent traits or abilities and thus D latent traits define a D-dimensional latent space, with each examinee's location in the latent space being determined by the examinee's position on each latent trait.

The difficulty of choosing between unidimensional and multidimensional item response models is manifest when information about how the test items were constructed is unavailable. Factor analysis is one way to check the assumption of unidimensionality. However, this approach has its own problems. For example, there might be a factor solution with too many factors resulting from using inappropriate correlation estimation such as the phi correlation or the tetrachoric correlation (McDonald & Ahlawat, 1974). When the knowledge of content domains specified by the test developer is available (usually from a blueprint of the items), a multidimensional item response model should be fitted to the test data if more than one ability is assumed to be measured. By conducting such an analysis, one can get information about how strongly the dimensions or the latent traits are correlated. It is often argued that when the traits are highly correlated, a unidimensional item response model can represent the data as well as a multidimensional model. But until we obtain such correlation information from the multidimensional analysis, blindly fitting a unidimensional item response model to multidimensional data sets can bias parameter estimation and person ability estimation (Folk & Green, 1989).

There are also situations when high correlations among dimensions are achieved, but factors other than the dimensionality of the traits being measured may have driven the correlations up. Item dependence could be one of the factors. The second basic assumption of item response modeling is local independence.

The assumption of local independence implies that given a person's ability, any response to one item is independent of the responses to other items. If we denote θ to be the latent person ability and x_i to be the observed response of the variable X_i for item i, the local independence assumption can be written as follows:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_I = x_I \mid \theta) = \prod_{i=1}^{I} P(X_i = x_i \mid \theta)\,. \tag{1}$$

This is, in fact, a very strong assumption. Many tests constructed for a short period of 40 to 45 minutes, and monitored in a classroom, do not necessarily meet this requirement.

For example, if a test is solely composed of short answer or multiple-choice questions following written stimulus materials, it can be difficult to provide different stimulus materials for each item with the time constraint. Therefore, one common practice is to have each piece of stimulus material followed by a small number of items. Obviously, the local independence assumption may be violated in this situation.

Many researchers have tried to address this dependence issue. Among them, Wainer and Kiely (1987) introduced the idea of *testlet* in computerized adaptive testing. The concept of testlet is to include the possibility of branching processes among items. Rosenbaum (1984, 1988) introduced the idea of *bundle independence* and this paper is going to follow this terminology, because it addresses the conditional independence issue. His idea is to create a bundle of the items that are expected to be dependent and to assume local independence across bundles. Suppose there is a set of C bundles and I_c is the number of items in each bundle; equation (1) is then modified in the following way:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_I = x_I \mid \theta) = \prod_{c=1}^{C} P(X_c = x_c \mid \theta), \text{ and } \sum I_c = I. \quad (2)$$

The distinction between the two equations is that x_i is an individual response on one item and x_c is a response pattern on a set of items in a bundle. In this sense, the number of response categories x_c can take will generally be larger than that x_i can. For instance, in a test that consists of all multiple-choice items, x_i can be 0 (incorrect) or 1 (correct). For a bundle of just two items, there can be four distinctive response patterns: (0 0), (1 0), (0 1), and (1 1). When the number of response categories in each item and the number of items in each bundle increase, the number of response patterns each bundle can possibly have will increase dramatically. A bundle of three polytomous items that each have 5 categories can have a maximum of 5^3=125 distinctive response patterns. This is why modeling interdependent items in a bundle can be complicated in terms of expressing the probability.

When a local dependence problem occurs in a set of test items that measure more than one latent ability, the interdependence of the items might cause the dimensions to be highly correlated, and thus multidimensional analysis will provide the misleading result that only one dimension needs to be modeled. By applying the concept of item bundles to multidimensional analysis, we can fit models that take account of multiple dimensions and item dependence simultaneously. The purpose of this paper is to investigate a dataset collected by the Science Education for Public Understanding Program (SEPUP) that has both dimensionality and dependence characteristics. An item bundle model nested in a multidimensional random coefficients multinomial logit (MRCML) model (Adams, Wang & Wilson, 1997) is applied to this data.

MODELS

The MRCML model is an extension of the unidimensional random coefficients multinomial logit (RCML) model (Adams & Wilson, 1996). The RCML model is a generalized Rasch model that provides the flexibility of customizing models for particular test situations. To keep the previously-used notation, θ is the latent variable and I is the total number of items, the probability of a response in category j of item i is modeled as

$$P(X_i = j; A, \bar{b}_i, \bar{\xi} \mid \theta) = \frac{\exp(b_{ij}\theta + \bar{a}'_{ij}\bar{\xi})}{\sum_{k=1}^{K_i} \exp(b_{ik}\theta + \bar{a}'_{ik}\bar{\xi})}, \tag{3}$$

where K_i = total number of response categories in item i,

$A = (\vec{a}'_{11},\ldots,\vec{a}'_{1K_1},\vec{a}'_{21},\ldots,\vec{a}'_{2K_2},\vec{a}'_{i1},\ldots,\vec{a}'_{iK_i})'$, a design matrix of *p* columns, relating observed responses to item parameters,

$\bar{b}_i = (b_{i1},\ldots,b_{iK_i})'$, a score vector of the response category from 1 to K_i for item i,

$\bar{\xi} = (\xi_1,\xi_2,\ldots,\xi_p)'$, a vector of *p* free item parameters,

$\vec{a}_{ik}$ = a design vector in matrix *A* for $i = 1, \ldots, I$; $k = 1, \ldots, K_i$.

The score vector b_i provides the flexibility of non one-to-one mapping between the category and the score that is allocated to that category. It can be collected into a large vector $\bar{b} = (\bar{b}_{11},\ldots,\bar{b}_{1K_1},\bar{b}_{21},\ldots,\bar{b}_{2K_2},\ldots,\bar{b}_{i1},\ldots,\bar{b}_{iK_i})'$, which allows different numbers of categories for different items. This provides the opportunity to calibrate both dichotomous and polytomous items at the same time. The vector of free parameters $\bar{\xi}$ and the design vector $\vec{a}_{ik}$, which is a linear combination of vector $\bar{\xi}$ determine how the model is specified. The vector $\bar{\xi}$ includes all the parameters that characterize the items, such as item difficulty, step difficulty, facet, interaction, etc. The design vector affords the possibilities of specifying customized models.

By extending the single latent variable to a *D*-dimensional latent space and collecting θ into a vector $\bar{\theta} = (\theta_1,\theta_2,\ldots,\theta_D)$, we can write the MRCML model as follows:

$$P(X_i = j; A, B_i, \bar{\xi} \mid \bar{\theta}) = \frac{\exp(\bar{b}'_{ij}\bar{\theta} + \vec{a}'_{ij}\bar{\xi})}{\sum_{k=1}^{K_i} \exp(\bar{b}'_{ik}\bar{\theta} + \vec{a}'_{ik}\bar{\xi})}. \tag{4}$$

The scoring vectors $\bar{b}_{ik} = (b_{ik1}, b_{ik2},\ldots,b_{ikD})'$ can be collected into a scoring sub-matrix $B_i = (\bar{b}'_{i1},\bar{b}'_{i2},\ldots,\bar{b}'_{iK_i})'$ for item *i* and furthermore into a larger scoring matrix

$B = (B_1', B_2', \ldots, B_I')'$ for the whole test. The distinction between equations (4) and (3) is that b_{ij} and θ are scalars in equation (3) whereas they are vectors in equation (4).

Now consider item bundles rather than individual items in the multidimensional case. Let K_c be the total number of distinctive response patterns in item bundle c. The probability of one response pattern j of bundle c can be modeled as

$$P(X_c = j; A, B_c, \bar{\xi} \mid \bar{\theta}) = \frac{\exp(\bar{b}_{cj}'\bar{\theta} + \bar{a}_{cj}'\bar{\xi})}{\sum_{k=1}^{K_c} \exp(\bar{b}_{ck}'\bar{\theta} + \bar{a}_{ck}'\bar{\xi})} . \tag{5}$$

As mentioned above, when the number of categories in the item and the number of items in the bundle increase, the denominator of equation (5) will be a long expression that takes into account all possible response patterns.

EXAMPLE

Data

The data set this paper will explore is from the field test of an assessment system for a yearlong middle school science curriculum, Issues, Evidence, and You (IEY) during the 1994-1995 school year. The curriculum was developed by the Science Education for Public Understanding Program (SEPUP). The assessment system includes SEPUP variables, assessment tasks, scoring guides, link tests and other components (Roberts, Wilson & Draney, 1997). There are five SEPUP variables that represent student learning corresponding to the core concepts of IEY. There are Designing and Conducting Investigation (DCI), Evidence and Tradeoffs (ET), Understanding Scientific Concepts (UC), Communicating Scientific Information (CSI) and Group Interaction (GI). Appendix A describes the SEPUP variables and sub-parts known as elements for each variable. The assessment tasks are performance assessments in which students were asked to produce a number of complex performances based on assessment activities. Scoring rubrics were established for each variable, listing the criteria for levels of student performance. The link tests are additional assessment activities for teachers to use at major course transitions that are also based on the SEPUP variables. During the field test year, data were collected for only four of the five SEPUP variables, because satisfactory methods for collecting data on the variable, Group Interaction, were not yet developed. It is the case in this field test data that a single piece of response was scored on multiple elements of a variable and even multiple variables. Teachers used the scoring guides to rate student performance into five ordered, qualitatively different categories, scored 0 through 4.

Previous analysis has been done on this data set by Draney and Peres (1998), investigating the multidimensional nature of the data, the change of student growth and rater severity over time. However, the problem of item dependence is still evident and needs to be examined.

Analysis I: Dimensionality

In the first set of analysis, a unidimensional model as well as a multidimensional model were fitted to the link test portion of the 1994-1995 field test data, using the (M)RCML program (Adams, Wilson & Wang, 1997). Appendix B shows the item number, the variable/element each item was supposed to measure in each link test and the linking item structure across three tests. As is apparent, most of the items were scored on multiple elements or variables. A subset of the observed scores was selected. The leftmost column in Appendix B indicates the numbering of the observed scores chosen for the analysis. Two factors were considered in the process of selecting these observed scores. The variable CSI was dropped from the analysis, since it was not assessed often. In order to carry out a multidimensional analysis, at least a reasonable number of observations should be obtained for each dimension. This resulted in three dimensions, DCI, ET and UC. In addition, a maximum of three scores were selected from a single item to keep the number of categories in each bundle manageable. In fact, after removing the variable CSI, only two items (LinkTest 2 item 1 and LinkTest 3 item 1) have more than three scores. They both have two scores on DCI, one on UC and one on ET. The decision was made to choose one of the two scores on DCI, as they measure the same element. An alternative approach could be taking the average of the two scores. This will involve a decision between rounding and truncating.

From now on, the observed scores will be referred to as items, and the original items in the link tests will be referred to as bundles labeled from L1i1 (LinkTest 1 item 1) to L3i3 (LinkTest 3 item 3). There are 22 items and 10 bundles in total.

As the link tests were given at three different times throughout the school year of 1994-1995, it seems reasonable to assume three different latent abilities on each latent dimension for an individual student. This is achieved by differentiating person *A* at time point 1 from person *A* at time point 2. Therefore, the data organization for this analysis includes three repetitions of 1383 students who took the link tests. The following chart shows how the data was organized. There are 4149 (=1383x3) rows (cases) and 22 columns (items)[1].

[1] The empty cells are systematic missing data.

ID	Responses		
1st 1383	Link Test 1		
2nd 1383		Link Test 2	
3rd 1383			Link Test 3

Model 1: Unidimensional Rating Scale Model

The data set consists of polytomous responses, thus, in addition to the item location parameters, the step difficulties of moving from one response category to a higher one in each item should be modeled. As items measuring one variable were scored according to the scoring rubric for that variable, it is assumed that the step difficulties do not vary across those items. This is the standard approach used in all SEPUP analysis. Therefore, a mixed rating scale model was fitted to the data. The item parameter vector $\bar{\xi}$ contains location parameters (δ) of 22 items and three sets of step parameters (τ), one set for each variable. Each variable has five score categories and the mean of the step parameters is constrained to zero, so only three step parameters are estimated per variable per test time; this yields a total of 31 (=22+3x3) parameters in vector $\bar{\xi}$. Item frequency statistics show that for almost half of the items, no student obtained the highest response, 4. In particular, these items are all from the bundles that have three items. The two items left in the three-item-bundles that have a few responses of 4 are item 7 and 22. These responses of 4 (about 0.01% of 1383 cases) were then recoded to missing so that within each bundle, items have an equal number of response categories. Thus, the scoring vector $\bar{b}$ contains a combined repetition of 0, 1, 2, 3, 4 and 0, 1, 2, 3. Appendix C shows the basic structure of the design matrix A. The estimates and standard errors of the parameters are listed in the second and third columns of Appendix D.

Model 2: Between-Item Multidimensional Rating Scale Model

Since the difference between the RCML and MRCML models only lies in the scoring matrix, a multidimensional mixed rating scale model was then fitted to the data using the same design matrix. The scoring matrix *B* has 3 columns (variables), indicating the dimension each item is supposed to measure. This is called a between-item multidimensional model (Adams, Wilson & Wang, 1997), because each item was only scored on one variable. There are again 31 parameters. In the fourth and fifth columns of Appendix D are the estimates and standard errors of these parameters. The following pairwise correlations of the three dimensions were calculated from the variance and covariance matrix (Appendix I) estimated by the (M)RCML program.

	DCI	ET	UC
DCI	1		
ET	0.73	1	
UC	0.76	0.82	1

Discussion I: Model 1 vs. Model 2

The correlations show that the three dimensions of latent ability are fairly closely related. The last column in Appendix D lists the differences of the parameter estimates from Model 1 and Model 2. They do not differ very much. It is still arguable whether multidimensional analysis is necessary because it might be the case that unidimensional analysis is sufficient. Therefore, a comparison of the fit statistics of the two models using the change in the likelihood ratio χ^2 was performed. Since the unidimensional rating scale model is a sub-model of the multidimensional rating scale model, a statistical significance test can be used to check which model fits the data better. The following table shows the deviance and degrees of freedom of the two models, and their differences.

	Model 1	Model 2	Difference
Deviance	28628.22	28522.67	105.55
DF	32 (=31+1)	37 (=31+3+3)	5

In both models, one assumption was made for the person ability distribution. It is assumed to have a normal distribution, with a mean of 0. Therefore, besides the item parameters, there is one more parameter, the variance of the person ability distribution, to be estimated in Model 1. In Model 2, since there are three dimensions, there are six more parameters to be estimated, three for the variances of the person ability distributions and three for the covariances between any two dimensions. The multidimensional model fits the data significantly better than the unidimensional one at α=0.01. This is consistent with the conclusion drawn by Draney and Peres on the entire SEPUP 1994-1995 field test data[2].

Analysis II: Dependence

Let us now investigate the item dependence issue. In order to perform the item bundle analysis, the data were recoded so that one score was given to each response pattern in a bundle. At most, three items were included in one bundle in order to make the recoding process manageable. For a bundle of two items, there are 5^2=25 categories coded from 0 to 24; and for a bundle of three items, there are 4^3=64 categories coded from 0 to 63. Appendix E shows the recoding schema[3].

<u>Model 3</u>*: Within-Item Multidimensional Item Bundle Model with Interactions*

This time the item parameter vector $\bar{\xi}$ contains some interaction parameters in addition to the regular location and step parameters. In a bundle of two items, 2-way

[2] Draney and Peres' study was carried out using the ConQuest program (Wu, Adams & Wilson, 1997).
[3] The cross-tab frequency statistics show that not every response category in the bundle has observations.

interactions (ω) take place when the examinee gets the same score on both items. In a bundle of three items, there are additional 3-way interactions (υ) when the examinee gets the same score on all three items. Among the 10 bundles, L2i2 and L2i5 just have one item in the bundle, and there are 4 bundles of two items and 4 bundles of three items. The first table in Appendix F shows a sub design matrix for a bundle that consists of two items. This is following the approach taken by Wilson and Adams (1995), and Hoskens and De Boeck (1997) to investigate local dependence characteristics. In particular, in Hoskens and De Boeck's study, they examined the interaction effects (ω) in addition to the main effects (δ) of the items. They did not model step parameters since their data set had only dichotomous items. In their analysis, the main effects are equivalent to the item location parameters in this analysis, and instead of modeling one location parameter per bundle, location parameters are still estimated on the item level. This is consistent with the recoding schema. However, the step parameters are estimated on the bundle level. Consider the *Partial Credit* model (Masters 1982) for two independent items with 5 response categories. After recoding the data following the 2-item-bundle recoding scheme, we will get a design matrix listed in the second table of Appendix F. The step parameters for the first and second items are indexed by $\tau_{1.}$ and $\tau_{2.}$ respectively. By comparing the two design matrices in Appendix F, we notice that the step parameters estimated on the bundle level are simply the additions of the corresponding step parameters estimated on the item level in the independent case. For instance, τ_1 in the first table is equivalent to the sum of τ_{11} and τ_{21} in the second table. In total, there are 3 steps for each two-item-bundle and one-item-bundle, and 2 steps for each three-item-bundle, because the step parameters are constrained to have a mean of 0. After adding the interaction parameters for each bundle, we finish modeling the dependence within a bundle. The interaction parameters indicate the additional difficulty (or easiness) of getting the same scores on two items, or three items in the bundle. The term "interaction" here is analogous to the interaction effect in ANOVA. This yielded a total of 60 (=22δ+(3x6+2x4)τ+8ω+4υ) item parameters. The scoring matrix *B* has 3 columns, with each column indicating one dimension. This time, it is a within-item multidimensional model, because some of the bundles were loaded on more than one dimension. Appendix G lists the parameter estimates and standard errors obtained from the (M)RCML program. The following correlation matrix was also calculated from the variance and covariance matrix (Appendix I).

	DCI	ET	UC
DCI	1		
ET	0.67	1	
UC	0.67	0.89	1

<u>Discussion II</u>*: Model 2 vs. Model 3*

The correlations obtained from the item bundle analysis differ a bit from those based on the multidimensional analysis. The correlation between DCI and ET, DCI and UC dropped while the correlation between ET and UC increased. By modeling the dependence between items, associations between the latent dimensions DCI and ET, DCI and UC were weakened. In fact, the skills required for designing and conducting investigations are quite different from using evidence to make tradeoffs and understanding concepts. The variables ET and UC belong to the domain of content knowledge in which the students are required to refer to the materials and concepts learned in the curriculum, whereas DCI belongs to the domain of process knowledge in which skills of "performing" are required. It is possible that the multidimensional item bundle analysis makes a better separation of the two knowledge domains because the dependence between these two have been taken into account. It is more difficult to explain why the correlation between ET and UC increased at this stage, though both ET and UC are content variables. Future work with simulated data sets can be done to investigate the likelihood of increased correlation within the same domain (content or process) and decreased correlation across the domains.

All the 2-way interaction parameter estimates from this model are negative. This implies that after modeling the dependence of the items within a bundle, items became easier than they were when the dependence was ignored. This additional easiness might be the evidence of the existence of item dependence. As teachers gave several scores on different variables/elements for a single piece of work, it is possible that the score he/she assigned on the second variable/element was affected by what he/she assigned on the first one. None of the 3-way interaction parameter estimates is statistically different from 0. This may suggest that modeling only 2-way interactions is sufficient for this data set.

Model 4: *Multidimensional Item Bundle Model without 3-way Interactions*

This model was fitted to the data based on the previous results. The vector $\bar{\xi}$ now has 56 parameters, after removing the four 3-way interaction parameters. The scoring matrix stays the same.

Model 5: *Multidimensional Item Bundle Model with All 2-way Interactions*

To further investigate the interaction effects on any two items in a bundle that has three items, a modified bundle model was fitted to the data by differentiating the 2-way interactions from each other within such bundles. In Model 3, only one 2-way interaction was used for each bundle containing three items. That was based on the hypothesis that getting the same score on any two of the three items may make a difference. However, this is not quite appropriate for the bundle that measures three different dimensions. For example, it is not necessarily the case that getting scores of 3 on both DCI and UC is the same as getting scores of 3 on both ET and UC. Therefore, three unique 2-way interactions were modeled for all 3-item-bundles. This resulted in a total of 64 parameters. Appendix H lists the estimates and standard errors.

<u>Discussion III</u>: Model 3 vs. Model 4 & Model 5

Similar to the multidimensional analysis, in addition to the item parameters, six parameters to describe the person distributions (3 variances and 3 covariances) need to be estimated in the item bundle analysis. The fit of Model 3, 4 and 5 are displayed as follows:

	Model 3	Model 4	Model 5
Deviance	26065.60	26053.43	26002.78
DF	66 (=60+3+3)	62 (=56+3+3)	70 (=64+3+3)

It is confirmed that 3-way interaction is not necessary for this data set, as Model 3 is not statistically better than Model 4. On the other hand, Model 5 in which 2-way interactions were differentiated in 3-item-bundles shows significant improvement in the deviance at $\alpha=0.01$.

The correlation matrices from Model 5 do not differ much from that of Model 3 (Appendix I). The item location and step estimates from these models are also similar to each other. Let us examine the 2-way interactions of Model 5 in detail. First of all, the estimates of the interactions are all negative, except for two. They are the interactions between items 5 and 6, and 13 and 14, however, these estimates are not significantly different from 0. In addition, in the last bundle L3i3, two of the three interaction estimates are insignificant. Even though there is no systematic pattern that which pairwise interaction always yields an easier (or more difficult) estimate, modeling three unique 2-way interactions reveals that getting the same score on some variables/elements may be easier than getting the same score on other variables/elements. There are two interactions that have relatively large negative values, bundle L1i2 and L1i4. These are the bundles measuring a single dimension. The dimensions they are measuring, UC and ET respectively, are both about content knowledge but are about different pieces of content knowledge. This might imply that for variables that are targeting content knowledge, teachers' ratings on one element are strongly influenced by their ratings on other elements.

CONCLUSION

The SEPUP 1994-1995 link tests have both multidimensionality and item dependence issues. Because the performance assessment was rather time-consuming, only a small number of items could be given to each student during the testing period. To make the most of these students' responses, each item was specifically designed to be multidimensional, and scored on a number of different variables/elements. Analysis that models only dimension or dependence alone is not adequate. A multidimensional item bundle analysis suggests that most of the items are dependent within each bundle, no matter whether the bundles are unidimensional or multidimensional. Interaction effects do exist for most of the pairwise score combinations. In particular, the effects are more prominent for the bundles in which items measure content knowledge. Taking the item dependence into account may make the correlation among latent dimensions more accurate and meaningful.

LIMITATION & SUGGESTION

Theoretically, violation of local independence can lead to inaccuracy of parameter estimation (Chen & Thissen, 1996) and of person proficiency estimation (Wilson, 1988). For item bundle analysis, estimating 3 dimensions using the quadrature method of numerical integration is extremely time consuming. Therefore, a relatively relaxed convergence criterion, 0.005, was used for the three bundle analyses compared to 0.001, which was used in unidimensional and multidimensional analysis. Therefore, it is hard to compare the accuracy of the parameter estimates obtained from Model 2 and Model 5. As for the interpretation of location and step estimates, they cannot be compared either, because the step parameters were estimated on a "variable" level in the multidimensional model whereas they were estimated on a "bundle" level in the item bundle model. Additionally, the data containing students' scores on each variable or element were recoded in the item bundle analysis. Due to this change in the data, the multidimensional model and the bundle models are not hierarchically ordered, so comparisons of the fit statistics of these models cannot be carried out using a likelihood ratio test.

Further examination of person ability distributions across dimensions and over time should be done for all the models. Fit statistics of items as well as of persons should also be checked. These results have not be obtained from the current analysis due to the limited output options in the (M)RCML program.

Another limitation of this analysis is that although test reliability can be calculated, it is not the one needed here, because not all the items were included in the analysis due to the fact that each item bundle used three items at most.

Future analysis can be conducted by creating models derived from model 5, adding more interaction parameters that characterize the difficulty of getting two or three scores on adjacent response categories. Individual response patterns show that it is rarely the case that someone is scored more than three categories apart on different variables or elements.

Finally, as mentioned before, simulation work on data that have both dimensionality and dependence features is worth pursuing.

REFERENCES

Adams, R.J., & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. Wilson (Eds.), *Objective Measurement III: Theory into practice*. Norwood, NJ: Ablex.

Adams, R. J., Wilson, M. & Wang, W. C. (1997). The Multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1-23.

Chen, W. H. & Thissen, D. (1996). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289.

Draney, K. & Peres, D. (1998). *Multidimensional modeling of complex science assessment data.* University of California, Berkeley: BEAR Report Series, SA-98-3.

Folk, V. G. & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement, 13*, 373-389.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods, 2*(3), 261-277.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

McDonald, R. P. & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27*, 82-99.

Roberts, L., Wilson, M. & Draney, K. (1997). *The SEPUP assessment system: an overview*. University of California, Berkeley: BEAR Report Series, SA-97-1.

Rosenbaum, P.R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425-435.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53*, 349-359.

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Wang, W. C. & Wilson, M. L. (1994). Comparing multiple choice and performance-based items using item response modeling. In G. Engelhard, Jr. & M. Wilson (Eds.), *Objective Measurement III: Theory into practice*. Norwood, NJ: Ablex.

Wang, W. C., Wilson, M. & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. In M. Wilson, G. Engelhard and K. Draney (Eds.), *Objective Measurement IV: Theory into practice*. Greenwich, CT: Ablex.

Wilson, M. (1988). Detecting and interpreting local item dependence using a family of Rasch models. *Applied Psychological Measurement, 12*, 353-364.

Wilson, M. & Adams, R.J. (1995). Rasch models for item bundles. *Psychometrika*, *60*(2), 181-198.

Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis, Rasch Measurement*, MESA Press, Chicago.

Wu, M. L., Adams, R. J. & Wilson, M. (1997). *ConQuest: Generalised item response modeling software, Draft Release 2*. Hawthorn, Australia: ACER.

Appendix A. SEPUP Variables and Elements

- Designing and Conducting Investigation (DCI): Designing a scientific experiment to answer a question or solve a problem, selecting appropriate laboratory procedures to collect data, accurately recording and logically displaying data (e.g. in graphs and tables), and analyzing and interpreting results of an experiment.
    1. designing investigation (di)
    2. selecting and recording procedures (srp)
    3. organizing data (od)
    4. analyzing and interpreting data (aid)

- Evidence and Tradeoffs (ET): Identifying objective, relevant scientific evidence, and evaluating the advantages and disadvantages of different possible solutions to a problem based on the evidence available.
    1. using evidence (ue)
    2. using evidence to make tradeoffs (uemt)

- Understanding Concepts (UC): Recognizing and applying relevant scientific concepts (e.g. threshold, measurement, properties of matter) to an investigation or problem solution.
    1. recognizing relevant content (rrc)
    2. applying relevant content (arc)

- Communicating Scientific Information (CSI): Organizing and presenting results, arguments, and conclusions in a way that is free of technical errors and effectively communicates with the chosen audience.
    1. organization (org)
    2. technical aspects (ta)

- Group Interaction (GI): Developing time management skills, the ability to work together with teammates to complete a task (such as a lab experiment) and to share the work of an activity.

Appendix B. Link Test Structure

Link Test 1

Number	Item	Variable	Element	Links back to number
1	1	DCI	di	
2	1	ET	uemt	
3	2	UC	rrc	
4	2	UC	arc	
5	3	DCI	di	
6	3	DCI	srp	
7	3	DCI	aid	
8	4	ET	ue	
9	4	ET	uemt	
	4	CSI	org	
	4	CSI	ta	
10	5	ET	ue	
11	5	ET	uemt	

Link Test 2

Number	Item	Variable	Element	Links back to number
	1	DCI	di	
12	1	UC	rrc	
13	1	DCI	di	
14	1	ET	uemt	
15	2	UC	arc	
	3	DCI	di	5
	3	DCI	srp	6
	3	DCI	aid	7
	4	CSI	org	
	4	CSI	ta	
	4	ET	ue	8
	4	ET	uemt	9
16	5	UC	arc	

Link Test 3

Number	Item	Variable	Element	Links back to number
	1	DCI	di	
17	1	UC	rrc	
18	1	DCI	di	
19	1	ET	uemt	
	2	UC	arc	16
20	3	DCI	srp	
21	3	DCI	od	
22	3	DCI	aid	
	4	ET	ue	10
	4	ET	uemt	11

Appendix C. Design Matrix *A* in Analysis I[4]

	Link Test 1 δ	Link Test 2 δ	Link Test 3 δ	DCI τ	ET τ	UC τ
Link Test1	0 0 0 -1 0 0 -2 0 0 -3 0 0 -4 0 0 0 0 0 0 -1 0 0 -2 0 0 -3 0 0 -4 0 0 0 0 0 0 -1 0 0 -2 0 0 -3 0 0 -4			0 0 0 -1 0 0 -1 -1 0 -1 -1 -1 0 0 0	0 0 0 -1 0 0 -1 -1 0 -1 -1 -1 0 0 0	0 0 0 -1 0 0 -1 -1 0 -1 -1 -1 0 0 0 ......
Link Test2		0 0 0 -1 0 0 -2 0 0 -3 0 0 0 0 0 0 -1 0 0 -2 0 0 -3 0 0 0 0 0 0 -1 0 0 -2 0 0 -3		0 0 0 -1 0 0 -1 -1 0 0 0 0	0 0 0 -1 0 0 -1 -1 0 0 0 0	0 0 0 -1 0 0 -1 -1 0 0 0 0 ......
Link Test3			0 0 0 -1 0 0 -2 0 0 -3 0 0 0 0 0 0 -1 0 0 -2 0 0 -3 0 0 0 0 0 0 -1 0 0 -2 0 0 -3	0 0 0 -1 0 0 -1 -1 0 0 0 0	0 0 0 -1 0 0 -1 -1 0 0 0 0	0 0 0 -1 0 0 -1 -1 0 0 0 0 ......

[4] For the purpose of visual clarity, repetitions of 0 are omitted in the table. The negative sign makes the interpretation of the parameters more meaningful. For instance, a low value of an item location parameter estimate means that item is relative easy.

Appendix D. Parameter Estimates from Analysis I

Par.	Description	Model 1 Unidimensional Estimate	S.E.	Model 2 Multidimensional Estimate	S.E.	□ of Est.
δ	1 (DCI)	0.218	0.033	0.225	0.033	-0.007
	2 (ET)	0.779	0.035	0.809	0.035	-0.030
	3 (UC)	0.211	0.029	0.242	0.028	-0.031
	4 (UC)	0.464	0.031	0.530	0.031	-0.066
	5 (DCI)	0.427	0.023	0.456	0.023	-0.029
	6 (DCI)	-0.203	0.018	-0.222	0.018	0.019
	7 (DCI)	0.544	0.024	0.581	0.024	-0.037
	8 (ET)	0.590	0.028	0.615	0.028	-0.025
	9 (ET)	1.181	0.032	1.240	0.032	-0.059
	10 (ET)	-0.238	0.025	-0.267	0.025	0.029
	11 (ET)	0.600	0.030	0.618	0.030	-0.018
	12 (UC)	-0.422	0.028	-0.471	0.028	0.049
	13 (DCI)	-0.077	0.026	-0.090	0.026	0.013
	14 (ET)	0.879	0.040	0.938	0.040	-0.059
	15 (UC)	0.435	0.030	0.491	0.030	-0.056
	16 (UC)	0.743	0.030	0.848	0.030	-0.105
	17 (UC)	0.404	0.049	0.484	0.049	-0.080
	18 (DCI)	0.232	0.042	0.245	0.042	-0.013
	19 (ET)	1.429	0.067	1.499	0.067	-0.070
	20 (DCI)	-0.104	0.037	-0.116	0.037	0.012
	21 (DCI)	-0.215	0.036	-0.236	0.036	0.021
	22 (DCI)	0.741	0.051	0.796	0.051	-0.055
τ	DCI 1	-0.293	0.035	-0.382	0.035	0.089
	DCI 2	-0.487	0.045	-0.492	0.045	0.005
	DCI 3	-0.859	0.125	-0.861	0.125	0.002
	ET 1	-1.341	0.040	-1.445	0.040	0.104
	ET 2	-0.981	0.047	-1.009	0.047	0.028
	ET 3	0.188	0.053	0.224	0.053	-0.036
	UC 1	-1.403	0.039	-1.622	0.039	0.219
	UC 2	-0.129	0.054	-0.170	0.054	0.041
	UC 3	0.291	0.076	0.361	0.076	-0.070

Appendix E. Data Recoding Table[5]

Two-Item-Bundle

		Item 1				
		0	1	2	3	4
	0	**0**	1	2	3	4
Item 2	1	5	**6**	7	8	9
	2	10	11	**12**	13	14
	3	15	16	17	**18**	19
	4	20	21	22	23	**24**

Three-Item-Bundle

Item 3 = 0

		Item 1			
		0	1	2	3
	0	*0*	1	2	3
Item 2	1	4	**5**	6	7
	2	8	9	**10**	11
	3	12	13	14	**15**

Item 3 = 1

		Item 1			
		0	1	2	3
	0	**16**	17	18	19
Item 2	1	20	*21*	22	23
	2	24	25	**26**	27
	3	28	29	30	**31**

Item 3 = 2

		Item 1			
		0	1	2	3
	0	**32**	33	34	35
Item 2	1	36	**37**	38	39
	2	41	41	*42*	43
	3	44	45	46	**47**

Item 3 = 3

		Item 1			
		0	1	2	3
	0	**48**	49	50	51
Item 2	1	52	**53**	54	55
	2	56	57	**58**	59
	3	60	61	62	*63*

[5] 2-way interactions are shown in bold numbers and 3-way interactions are shown in italic numbers.

Appendix F.

Design Matrix for a Two-Item-Bundle in Analysis II

δ_1	δ_2	τ_1	τ_2	τ_3	ω
0	0	0	0	0	-1
-1	0	-1	0	0	0
-2	0	-1	-1	0	0
-3	0	-1	-1	-1	0
-4	0	0	0	0	0
0	-1	-1	0	0	0
-1	-1	-2	0	0	-1
-2	-1	-2	-1	0	0
-3	-1	-2	-1	-1	0
-4	-1	-1	0	0	0
0	-2	-1	-1	0	0
-1	-2	-2	-1	0	0
-2	-2	-2	-2	0	-1
-3	-2	-2	-2	-1	0
-4	-2	-1	-1	0	0
0	-3	-1	-1	-1	0
-1	-3	-2	-1	-1	0
-2	-3	-2	-2	-1	0
-3	-3	-2	-2	-2	-1
-4	-3	-1	-1	-1	0
0	-4	0	0	0	0
-1	-4	-1	0	0	0
-2	-4	-1	-1	0	0
-3	-4	-1	-1	-1	0
-4	-4	0	0	0	-1

Design Matrix for Two Independent Items with 5 Response Categories

δ_1	δ_2	τ_{11}	τ_{12}	τ_{13}	τ_{21}	τ_{22}	τ_{23}
0	0	0	0	0	0	0	0
-1	0	-1	0	0	0	0	0
-2	0	-1	-1	0	0	0	0
-3	0	-1	-1	-1	0	0	0
-4	0	0	0	0	0	0	0
0	-1	0	0	0	-1	0	0
-1	-1	-1	0	0	-1	0	0
-2	-1	-1	-1	0	-1	0	0
-3	-1	-1	-1	-1	-1	0	0
-4	-1	0	0	0	-1	0	0
0	-2	0	0	0	-1	-1	0
-1	-2	-1	0	0	-1	-1	0
-2	-2	-1	-1	0	-1	-1	0
-3	-2	-1	-1	-1	-1	-1	0
-4	-2	0	0	0	-1	-1	0
0	-3	0	0	0	-1	-1	-1
-1	-3	-1	0	0	-1	-1	-1
-2	-3	-1	-1	0	-1	-1	-1
-3	-3	-1	-1	-1	-1	-1	-1
-4	-3	0	0	0	-1	-1	-1
0	-4	0	0	0	0	0	0
-1	-4	-1	0	0	0	0	0
-2	-4	-1	-1	0	0	0	0
-3	-4	-1	-1	-1	0	0	0
-4	-4	0	0	0	0	0	0

Appendix G. Parameter Estimates from Model 3

Par.	Description	Estimate	S.E.
δ	1 (DCI)	0.538	0.094
	2 (ET)	1.207	0.098
	3 (UC)	0.142	0.071
	4 (UC)	1.107	0.086
	5 (DCI)	0.433	0.037
	6 (DCI)	-0.298	0.032
	7 (DCI)	0.566	0.037
	8 (ET)	0.143	0.051
	9 (ET)	1.222	0.065
	10 (ET)	-0.140	0.076
	11 (ET)	1.295	0.087
	12 (UC)	-0.374	0.044
	13 (DCI)	-0.207	0.046
	14 (ET)	0.425	0.050
	15 (UC)	0.643	0.069
	16 (UC)	1.008	0.069
	17 (UC)	0.318	0.079
	18 (DCI)	0.116	0.077
	19 (ET)	1.030	0.089
	20 (DCI)	-0.134	0.067
	21 (DCI)	-0.243	0.066
	22 (DCI)	0.709	0.078
τ	L1i1 1	-1.637	0.110
	L1i1 2	-1.049	0.126
	L1i1 3	-0.610	0.122
	L1i2 1	-2.376	0.068
	L1i2 2	-0.572	0.077
	L1i2 3	0.618	0.090
	L1i3 1	-0.230	0.042
	L1i3 2	-0.505	0.052
	L1i4 1	-1.980	0.058
	L1i4 2	-0.720	0.063
	L1i4 3	0.460	0.073
	L1i5 1	-2.258	0.094
	L1i5 2	-1.680	0.105
	L1i5 3	0.097	0.092
	L2i1 1	-0.433	0.054
	L2i1 2	0.027	0.075
	L2i2 1	-1.976	0.110
	L2i2 2	-0.776	0.139
	L2i2 3	0.454	0.159
	L2i5 1	-2.179	0.094
	L2i5 2	-0.373	0.125
	L2i5 3	0.396	0.157
	L3i1 1	-0.420	0.087
	L3i1 2	-0.249	0.115
	L3i3 1	-0.059	0.093
	L3i3 2	-0.372	0.120

Par.	Description[6]	Estimate	S.E.
ω	L1i1	-0.469	0.099
	L1i2	-1.704	0.104
	L1i3	-0.463	0.069
	L1i4	-1.551	0.073
	L1i5	-0.549	0.095
	L2i1	-0.397	0.101
	L3i1	-0.413	0.152
	L3i3*	-0.208	0.143
υ	L1i3*	0.268	0.194
	L2i1*	0.042	0.274
	L3i1*	-0.205	0.433
	L3i3*	-0.165	0.414

[6] An asterisk indicates that the parameter estimate is not statistically different from 0.

Appendix H. Parameter Estimates from Model 5

Par.	Description	Estimate	S.E.
δ	1 (DCI)	0.529	0.093
	2 (ET)	1.192	0.098
	3 (UC)	0.133	0.070
	4 (UC)	1.081	0.086
	5 (DCI)	0.364	0.038
	6 (DCI)	-0.271	0.032
	7 (DCI)	0.561	0.040
	8 (ET)	0.136	0.051
	9 (ET)	1.212	0.065
	10 (ET)	-0.149	0.075
	11 (ET)	1.283	0.087
	12 (UC)	-0.389	0.046
	13 (DCI)	-0.139	0.048
	14 (ET)	0.386	0.050
	15 (UC)	0.629	0.068
	16 (UC)	0.987	0.068
	17 (UC)	0.324	0.081
	18 (DCI)	0.068	0.080
	19 (ET)	1.030	0.091
	20 (DCI)	-0.109	0.070
	21 (DCI)	-0.304	0.069
	22 (DCI)	0.727	0.080
τ	L1i1 1	-1.626	0.110
	L1i1 2	-1.041	0.126
	L1i1 3	-0.608	0.121
	L1i2 1	-2.332	0.068
	L1i2 2	-0.560	0.076
	L1i2 3	0.594	0.090
	L1i3 1	-0.225	0.040
	L1i3 2	-0.501	0.052
	L1i4 1	-1.980	0.058
	L1i4 2	-0.719	0.063
	L1i4 3	0.461	0.073
	L1i5 1	-2.253	0.093
	L1i5 2	-1.674	0.105
	L1i5 3	0.099	0.092
	L2i1 1	-0.403	0.051
	L2i1 2	0.021	0.074
	L2i2 1	-1.953	0.110
	L2i2 2	-0.767	0.139
	L2i2 3	0.445	0.158
	L2i5 1	-2.151	0.093
	L2i5 2	-0.365	0.125
	L2i5 3	0.385	0.156
	L3i1 1	-0.416	0.083
	L3i1 2	-0.250	0.114
	L3i3 1	-0.058	0.089
	L3i3 2	-0.369	0.119

Par.	Description[7]	Estimate	S.E.
ω	L1i1	-0.471	0.099
	L1i2	-1.714	0.103
	L1i3(5/6)*	0.000	0.079
	L1i3(6/7)	-0.467	0.074
	L1i3(5/7)	-0.639	0.065
	L1i4	-1.549	0.073
	L1i5	-0.549	0.094
	L2i1(12/13)	-0.732	0.097
	L2i1(13/14)*	0.102	0.109
	L2i1(12/14)	-0.490	0.105
	L3i1(17/18)	-0.375	0.165
	L3i1(18/19)	-0.727	0.155
	L3i1(17/19)	-0.420	0.163
	L3i3(20/21)*	-0.213	0.152
	L3i3(21/22)	-0.568	0.155
	L3i3(20/22)*	-0.075	0.162

[7] An asterisk indicates that the parameter estimate is not statistically different from 0.

Appendix I. Variance/Covariance Matrices from Multidimensional Models[8]

Model 2

	DCI	ET	UC
DCI	0.65240		
ET	0.45983	0.61601	
UC	0.55000	0.57838	0.80853

Model 3

	DCI	ET	UC
DCI	0.46700		
ET	0.37871	0.69195	
UC	0.49650	0.80563	1.19231

Model 4

	DCI	ET	UC
DCI	0.46686		
ET	0.38217	0.68669	
UC	0.49856	0.81916	1.17241

Model 5

	DCI	ET	UC
DCI	0.45131		
ET	0.36944	0.69704	
UC	0.48287	0.80756	1.17209

[8] The diagonal values are variances and the off-diagonal values are covariances.

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 9/97)